Title
propnet: A Knowledge Graph for Materials Science

Journal
Matter, 2(2)

ISSN
2590-2393

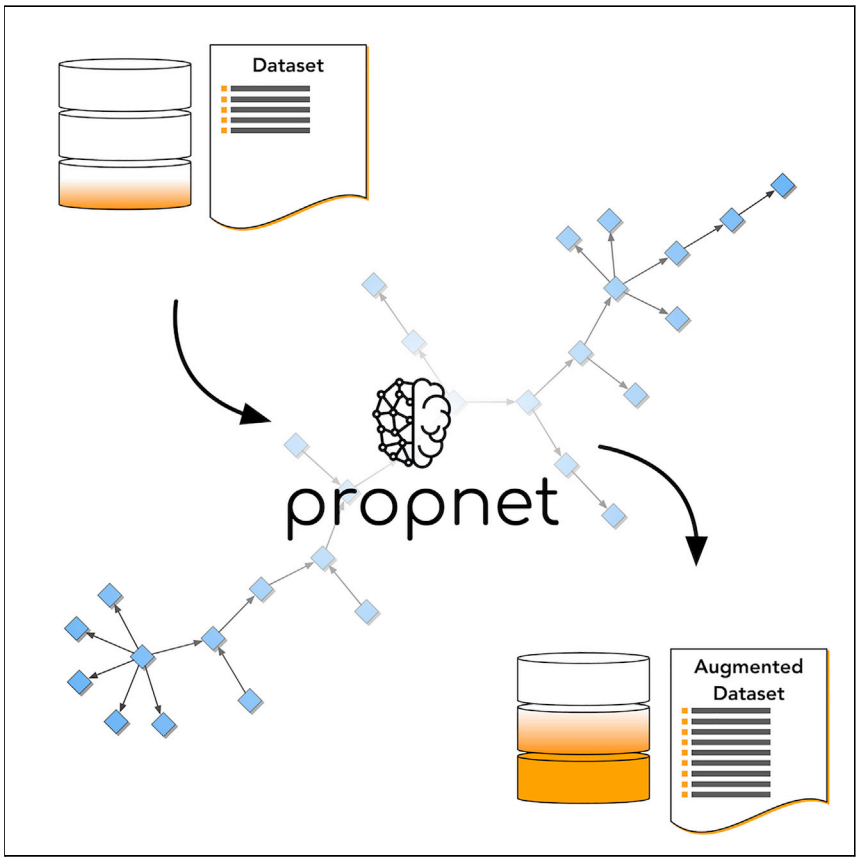
Authors
Mrdjenovich, D
Horton, MK
Montoya, JH
et al.

Publication Date
2020-02-05

DOI
10.1016/j.matt.2019.11.013

Peer reviewed

propnet: A Knowledge Graph for Materials Science



propnet is a computational framework to explore the network of relationships between fundamental materials properties. There exist many equations and models known from the materials science literature that provide the links between these properties, and this allows the representation of property connections as a larger, interconnected graph. Exploring this graph in a systematic away allows the automatic augmentation of existing materials databases and also provides new ways to gain insight into the relationships between the material properties themselves.

David Mrdjenovich, Matthew K. Horton, Joseph H. Montoya, ..., Vahe Tshitoyan, Anubhav Jain, Kristin A. Persson

kapersson@lbl.gov

HIGHLIGHTS

The relationships between properties of materials can be represented as a graph

Additional properties can be calculated automatically from an initial dataset

Multiple routes to calculate the same property can be evaluated to assess uncertainty

Available as an open-source Python code, propnet, and interactive website

propnet: A Knowledge Graph for Materials Science

David Mrdjenovich,[1,4] Matthew K. Horton,[2,4] Joseph H. Montoya,[3,4] Christian M. Legaspi,[2] Shyam Dwaraknath,[2] Vahe Tshitoyan,[2] Anubhav Jain,[2] and Kristin A. Persson[1,2,5,*]

SUMMARY

Data-driven materials science is bolstered by the recent growth of online materials databases. However, the current informatics infrastructure has yet to unlock the full knowledge available within existing datasets or to explore connections between different materials science domains. Here, we present a streamlined system for codifying and connecting materials properties in an open-source Python framework: `propnet`. We demonstrate the capability of this framework to augment existing datasets of materials properties: by consecutively applying a network of physical relationships to calculate related information, propnet connects disparate domain knowledge. Beyond an immediate increase in available information, the results allow for the examination of corre-lations between sets of properties and guide the design of multifunctional ma-terials. By emphasizing code extensibility and simplicity, we offer this software to the materials science community for general application to any experimental or computationally derived materials database.

INTRODUCTION

The field of science is one of daunting breadth, divided into numerous subfields and intersectional knowledge domains. While any one textbook may offer a basic over-view of the field or explore a particular niche, no single resource exists that can comprehensively chronicle centuries of established knowledge. Even in subfields of science, such as materials science, knowledge and discoveries are scattered over disparate publications and only known by experts of that particular specializa-tion. This lack of connection is an inherent challenge to the advancement of modern science, affecting multifunctional materials in particular. There is an ever-increasing demand for such materials, which consistently outperform traditional solutions, thus reducing size, weight, cost, power consumption, and complexity.[1]

In recent years, the field of materials informatics has blossomed, fueled by the growth of free online computed and experimentally derived materials data-bases.[2–9] However, at present these resources are unconnected and typically display only directly computed or measured data: few physically related properties are accessible. This is a notable opportunity loss, as materials properties are inherently interconnected. For example, the electronic structure of a material relates to its chemistry and geometry, which affects its energy-absorption capabilities, its refractive index, and its dielectric breakdown strength.

From a holistic perspective, materials science knowledge can be described as a network of relationships. By leveraging the organizational structure of these connections at scale, it is possible to gain new insight previously hidden within existing

Progress and Potential

Discovering the ideal material for a new application involves determining its numerous properties, such as electronic, mechanical, or thermodynamic, to match those of its desired application. The rise of high-throughput computation has meant that large databases of material properties are now accessible to scientists. However, these databases contain far more information than might appear at first glance, since many relationships exist in the materials science literature to derive, or at least approximate, additional properties.

propnet is a new computational framework designed to help scientists to automatically calculate additional information from their datasets. It does this by constructing a network graph of relationships between different materials properties and traversing this graph. Initially, propnet contains a catalog of over 100 property relationships but the hope is for this to expand significantly in the future, and contributions from the community are welcomed.

datasets. Specifically, we anticipate an increase in derived materials properties, a facilitated examination of property relationships, an improved uncertainty quantification between different models, and the ability to infer previously unknown physical correlations. By encoding canonical and novel materials property relationships in a facile serialized format (e.g., YAML format), software can predict physical behavior automatically, thus unlocking an array of latent information. Similarly, this process can be used to augment training sets with physically relevant property descriptors, improving the efficacy of predictive machine-learning models.

To realize this vision, we introduce propnet, an open-source Python package, as a means to programmatically codify and apply any facet of materials science knowledge. At its core, propnet is a growing catalog of materials properties, appropriate units, and property relationships, stored in extendable, general formats. For example, propnet stores atomic density (ρ_{atom}, in atoms/Å^3) and mean sound velocity (v_m, in m/s) as properties that connect to Debye temperature (T_D, in K) via the Debye model:

$$T_D = \frac{\hbar}{k_b}\left(6\pi^2 \rho_{atom}\right)^{1/3} v_m. \qquad \text{(Equation 1)}$$

Beyond simple equations and fundamental properties, propnet can store any property or relationship that can be expressed programmatically in Python, including varied properties such as crystal dimensionality, material cost, and other relationships with complex manipulation of inputs. As of propnet v.2019.07, the catalog consists of 115 materials properties and 69 relationships, or "models."

Together, these properties and relationships form a directed graph data structure (Figure 1) capable of representing arbitrarily complex property relationships, including those which are uni- and bidirectional. The utility of `propnet` lies in deriving an augmented set of materials properties from inputs provided by applying its graph traversal algorithm. For example, some databases report a computed band gap but do not use this band gap to estimate refractive index. Using `propnet`, the estimated refracted index is reported automatically. In this article, we demonstrate our feature-complete infrastructure, which, when applied to the Materials Project database, produces an average of 29 new properties per material, a 9-fold increase in available property values over the original data.

In addition to the core functionality of expanding datasets, we envision `propnet` as a valuable resource for materials informatics. Specifically, the datasets generated by `propnet` are not only inherently useful but also are suitable for generating physically motivated feature vectors. Such vectors are important for improving machine-learned models in materials design.[10] Using `propnet`, it is possible to assess the accuracy of property relationships as well as create ensembles of physical models that outperform any single model. In the following, we explore two examples of this more advanced functionality.

RESULTS AND DISCUSSION

As an initial proof of concept, `propnet` was applied to the Materials Project database, a publicly available, rapidly growing repository of computed materials properties, containing over 120,000 different materials.[4] The Materials Project[4] is a database mainly derived from first principles, affording each entry a minimum of four basic properties: the lattice and basis, the calculated band gap, and the total energy, as computed by density functional theory (DFT)[11–14] using the Perdew-Burke-Ernzerhof (PBE)[15] functional or PBE with Hubbard U[16] correction. In addition,

[1]Department of Materials Science & Engineering, University of California, 210 Hearst Mining Building, Berkeley, CA 94720, USA
[2]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[3]Toyota Research Institute, 4440 EL Camino Real, Los Altos, CA 94022, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: kapersson@lbl.gov

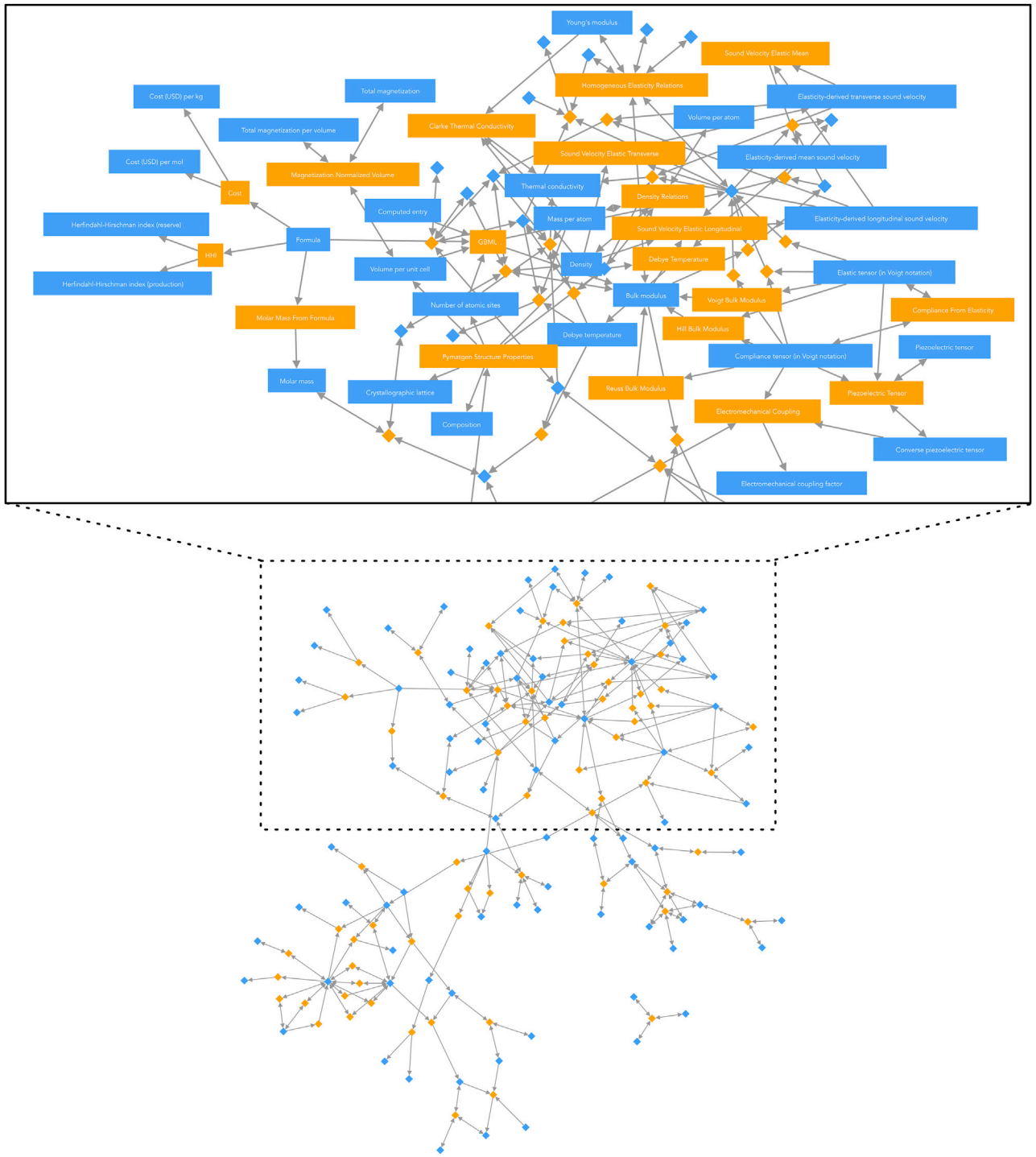


Figure 1. The Propnet Graph

Representation of the `propnet` knowledge graph, showing the connections between properties (blue) via property relationships or models (orange). Lines between shapes indicate that the connected model takes the connected property as input or yields the connected property as output, as represented by arrowheads pointing toward models or toward properties, respectively. The top inset shows a closer view of one section of the graph, including labels for selected models and properties.

the Materials Project provides several other structure-related properties, including chemical formula, atomic density, mass density, volume of the unit cell, and volume per atom. Together, these properties will be referred to as "base properties" for the remainder of the paper. For a growing subset of materials, currently numbering between 1,000 and 15,000, dielectric, elastic, piezoelectric, and vibrational properties have also been calculated alongside surface energies.[17–21] These tensor properties contain rich information and are used by `propnet` to substantially increase the amount of information available for each material. In the following, we use the Materials Project to demonstrate the capabilities of `propnet`; however, we emphasize that `propnet` is inherently agnostic to the data's origin and can be used universally on any experimental or computational materials property dataset.

Demonstrating the Increase in Derived Properties

To assess `propnet`'s effectiveness in expanding datasets, we chose an example material: wurtzite CdTe (space group: 186, Materials Project ID: mp-12779). CdTe is a semiconductor used in high-efficiency, thin-film photovoltaics, noteworthy for its optimal narrow band gap and ease of manufacture.[22] In the Materials Project, CdTe presents calculated dielectric and elastic tensors in addition to its base properties. After importing a dataset of 20 values for 20 distinct properties from Materials Project and executing `propnet`, a total of 629 new distinct property values spanning 41 new properties are derived. Each derived value represents a unique chain of physical relationships used to calculate a property that was previously hidden.

In a further demonstration, we illustrate the performance of `propnet` on datasets with variable data by systematically removing property values from the original Materials Project CdTe dataset. To this end, all possible subsets of data were generated, starting from ten initial materials properties. The reduced datasets include ten subsets of nine total properties, 45 subsets of eight total properties, and so forth. Every subset is then inspected by `propnet`, resulting in varying numbers of output property values for each subset. Grouping the data by subset size, we plot the mean number of total property values (derived + starting property values) as a function of subset size, demonstrating the ability of `propnet` to augment even very small datasets (Figure 2).

Similarly, we highlight the increased effectiveness of `propnet` when presented with diverse information. Starting with eight total values for the eight base properties of CdTe, we apply `propnet`, resulting in 21 new property values spanning 16 new properties. Next, we expand the initial dataset, first adding the dielectric tensor and then adding both the dielectric and elastic tensors. The addition of these tensors greatly increases the number of new properties: 21 and 38 are derived, respectively. The total number of property values also increases by factors of 5.5 and 35, respectively (Figure 2) relative to the initial eight data points. The elastic tensor substantially augments the dataset due to the numerous elasticity-based models currently available in `propnet`. In particular, some derived properties from the elastic tensor, such as bulk modulus, can have upper and lower bound values depending on the specific assumptions used (e.g., constant strain or constant stress). These many options then generate inputs for other models, leading to a compounding effect.

Identifying Correlations in Materials Properties

To analyze `propnet`'s potential for revealing relationships between materials properties, we use our framework to inspect each material entry in the Materials Project. Every entry includes the aforementioned "base properties" (lattice, basis, chemical formula, band gap, total energy, atomic density, mass density, volume of the unit

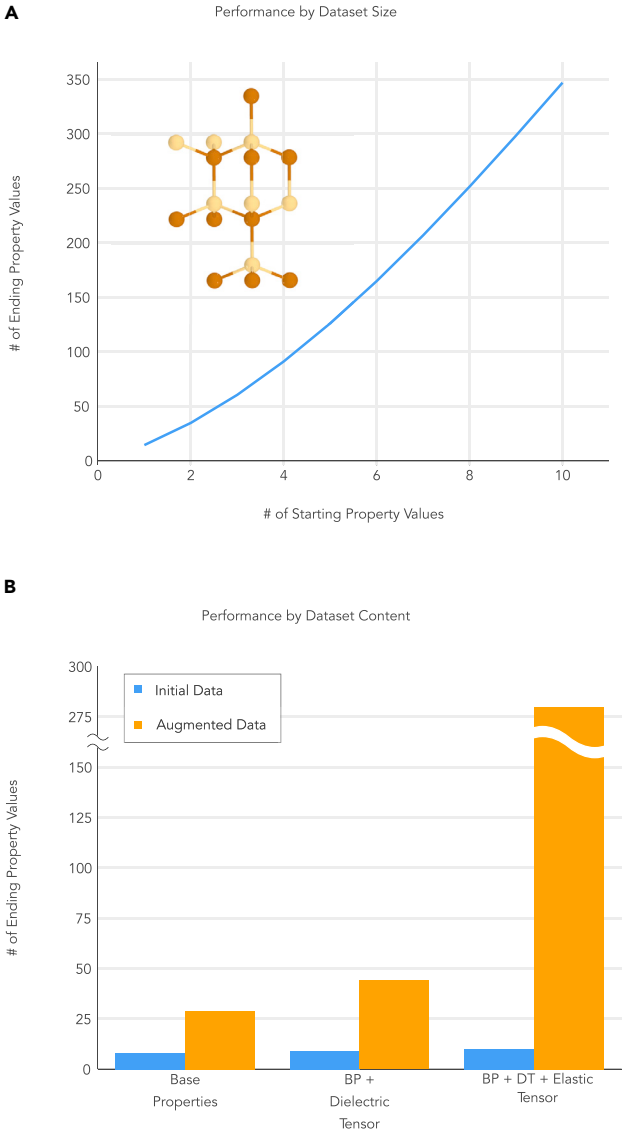


Figure 2. Augmenting a Dataset Using Propnet
The ability of `propnet` to augment materials science datasets by varying the (A) dataset size and the (B) dataset content using Materials Project data on wurtzite CdTe. In (A), the x axis represents the number of property values in the input set. The y axis represents the mean number of total (derived + initial) distinct property values, averaged over all possible input sets of size *x* after `propnet` is applied to the dataset. Note that `propnet` may derive several different property values for a given materials property. At top left is a rendering of the wurtzite CdTe crystal structure. In (B), the x axis represents the type of properties included as input to `propnet` and the y axis represents the number of resulting property values.

cell, and volume per atom). Elastic, dielectric, and piezoelectric tensors are included in the initial properties if they are available, along with a characterization of the magnetic ordering. Starting with these values, the datasets of each material are expanded using `propnet`, yielding up to approximately 30 distinct scalar properties.

To determine the degree of correlation between scalar properties, the mean value for each is collected material by material. For a given pair of scalar properties, the

materials that contain data for both are selected. The correlation is then measured using the maximal information coefficient (MIC) score.[23] This metric is robust to outliers and is capable of identifying nonlinear relationships. The MIC score ranges from 0 to 1, with 0 indicating no relationship between two variables and 1 representing a strong, monotonic relationship. When provided with a dataset, `propnet` can perform this correlation analysis automatically using the MIC score or a variety of other correlation metrics, including Pearson correlation, Spearman rank correlation, and Theil-Sen regression.

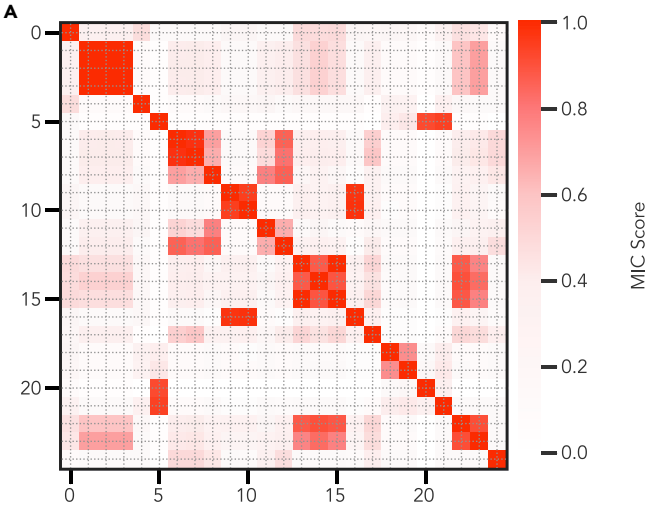
The correlation analysis (Figure 3) identifies 12 strong relationships with MIC scores above 0.9 (Figure 3). For example, and not unexpectedly, we find that the Debye temperature and the longitudinal sound velocity are related, with a MIC score of 0.905. This occurs because of the direct relationship between the Debye temperature and the mean sound velocity that is codified within a `propnet` model (see Equation 1). Noting that the Debye temperature is also proportional to the atomic density, the additional relationship causes variation that lowers the MIC score.

As a measure of property connectivity, a "graphical distance" can be used to highlight correlations between properties that are unexpected. We define the graphical distance between two properties to be the size of the smallest input set that derives both properties. Thus, the larger the graphical distance, the "harder" it is to connect those properties on the knowledge graph: properties that are correlated yet separated by a large distance in the `propnet` graph suggest that a model exists connecting the properties. If either of the properties is not connected to the `propnet` knowledge graph by any models, the graphical distance is considered undefined, and the properties are considered "not connected." For example, from an input set of total mass of a unit cell (m), unit cell volume (V), and number of atomic sites in a unit cell (n), we can derive the density ($\rho = m/V$) and the atomic density ($\rho_{\text{atom}} = n/V$). Density and atomic density both require two inputs to be derived from this input set but have a graphical distance of 3. The set of three variables (m, n, and V) derives both densities.

While heuristic, this metric gauges the conceptual separation between two properties in the domain of scientific knowledge. The models in `propnet` represent property relationships as they are presented in the literature, with the most atomic equations chosen. As such, this definition represents how a scientist might logically draw a path between two properties. We recognize that graphical distance can be defined in a number of ways, and we plan to explore different definitions as the knowledge graph grows.

As expected, many properties that are related to a high MIC score have a low graphical distance. Exploring this observation, we enumerate all property pairs and group them by graphical distance (Figure 4). Interestingly, some property pairs exhibit low correlation scores, even at low graphical distance. The definition of graphical distance assumes that the more information that is required to derive the two properties, the less correlated they are likely to be. This metric also assumes that all properties hold a relatively similar amount of information. However, properties such as tensor values or the crystal structure contain much more information than scalar values. If two properties are both derived from a single property such as these, they will have a graphical distance of 1 unit. Many of the observed low-score, low-distance property pairs are derived directly from these complex properties, affording them an artificially short graphical distance. This highlights a shortcoming of the definition for graphical distance.

A



0. Mean Mass Per Atom in Unit Cell
1. Volume Per Atom in Unit Cell
2. Atomic Density
3. Average Interatomic Spacing
4. Mass Density
5. Band Gap, Estimated Correction from Simulation
6. Young's Modulus
7. Shear Modulus
8. Bulk Modulus
9. Pugh Ratio
10. Poisson Ratio
11. Lame First Parameter
12. Pressure Wave Modulus
13. Transverse Sound Velocity
14. Longitudinal Sound Velocity
15. Mean Sound Velocity
16. Gruneisen Parameter
17. Vickers Hardness
18. Relative Permittivity
19. Relative Permeability
20. Refractive Index
21. Polycrystalline Optical Dielectric Constant
22. Debye Temperature
23. Thermal Conductivity
24. Melting Point

B

Property Pairs with MIC Score > 0.9

Rank	Property Pair	Graphical Distance (# of properties)	MIC Score
1	Atomic Density / Average Interatomic Spacing	2	1
2	Volume Per Atom in Unit Cell / Atomic Density	2	1
3	Average Interatomic Spacing / Volume Per Atom in Unit Cell	2	1
4	Mean Sound Velocity / Transverse Sound Velocity	7	1
5	Poisson Ratio / Gruneisen Parameter	7	0.979
6	Shear Modulus / Young's Modulus	3	0.975
7	Gruneisen Parameter / Pugh Ratio	7	0.967
8	Poisson Ratio / Pugh Ratio	3	0.945
9	Band Gap, Estimated Correction from Simulation / Polycrystalline Optical Dielectric Constant	Not Connected	0.941
10	Band Gap, Estimated Correction from Simulation / Refractive Index	1	0.919
11	Debye Temperature / Thermal Conductivity	9	0.913
12	Longitudinal Sound Velocity / Debye Temperature	9	0.905

Figure 3. Materials Property Correlations

(A) A correlation matrix heatmap demonstrating relationships between selected scalar materials properties derived from Materials Project data. Each numbered row/column corresponds to the numbered property in the list on the right. The color depth at each line intersection corresponds to the maximal information coefficient (MIC) score for the pair of properties intersecting at that point on the heatmap.

(B) A table of the scalar property pairs shown in the matrix with MIC scores >0.9, their graphical distance in number of properties (or "Not connected" for unconnected properties), and their MIC scores ranked in descending order.

We also observe a number of property pairs with high correlation scores that are deemed "not connected" on the `propnet` graph. These pairs suggest the presence of additional models that are missing from `propnet`. For example, the band gap

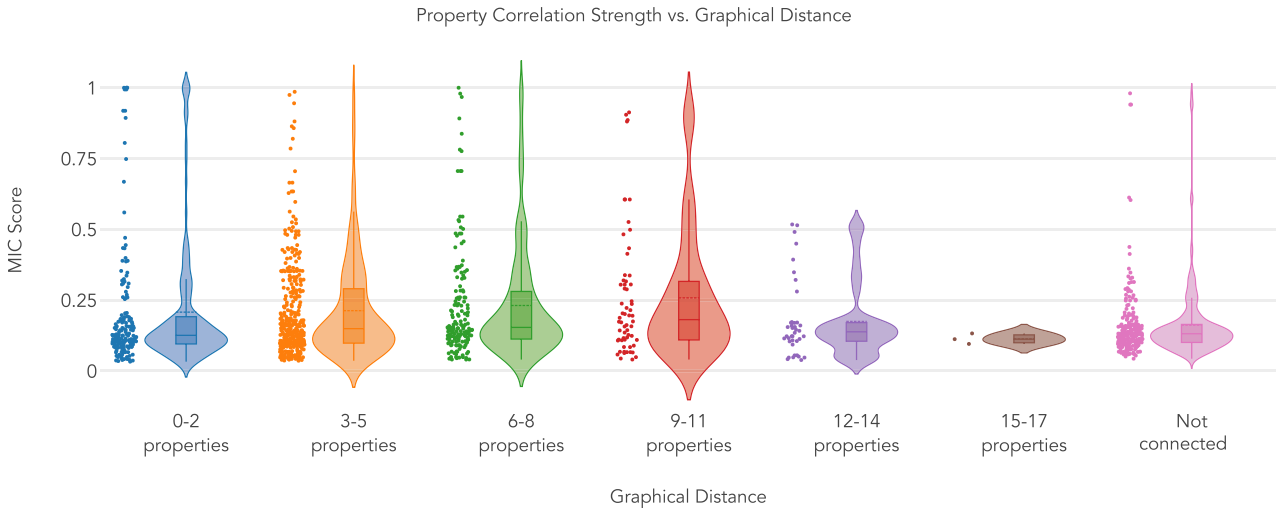


Figure 4. Property Correlations by Graphical Distance

Correlation strength between pairs of scalar properties grouped by their graphical distance into "violin" plots. The x axis represents the graphical distance separating two properties on the `propnet` graph, as defined in the text. "Not connected" means that the two properties are not connected by any property relationships to the current graph. The y axis represents the MIC score calculated for each pair of properties of a given graphical distance. The violin plot consists of a kernel density plot overlaid with a standard box-and-whisker plot. The kernel density plots correspond to the density of points at a given MIC score value for a given graphical distance. In the box-and-whisker plots, the solid horizontal lines represent the median and quartile divisions of the MIC scores and the dashed horizontal line represents the mean MIC score. To the left of each violin plot is a scatterplot of the raw MIC scores for each property pair with a given graphical distance value.

and the polycrystalline optical dielectric constant have a high correlation score (0.941) but are not connected. However, the relationship between the band gap, the refractive index, and the dielectric constant has been studied extensively.[24–29] As such, by examination of these correlation scores, `propnet` reveals knowledge missing from its library.

We see a marked decrease in the number of property pairs with high correlation as graphical distance increases. Hence, we suggest that `propnet` can be used to reveal analytic relationships, known or yet unknown, between properties by heuristically analyzing large MIC score outliers at larger graphical distances. As a case study, we consider the correlation between Debye temperature and Clarke thermal conductivity.

Of the top seven properties that correlate strongly with Clarke thermal conductivity, Debye temperature ranks the highest, with an MIC score of approximately 0.913 (Figure 5). This is particularly compelling because the two properties are separated by 9 graphical distance units on the current `propnet` knowledge graph (Figure 5). At this distance, most properties exhibit low correlation scores; however, the high correlation can be explained by exploring how these two properties are derived. The Clarke thermal conductivity emerges from considering the high-temperature limit of the Debye model,[30] yielding a lower bound estimate of the actual thermal conductivity. Taking the sound velocity (v_s) to be roughly proportional to the square root of Young's modulus (Y), $v_s \approx \sqrt{Y/\rho}$, it then follows that there exists a simple relationship connecting the Clarke thermal conductivity (κ_C) and the Debye temperature (T_D):

$$\frac{\kappa_C}{T_D} \approx \frac{(0.87)k_B^2}{\pi\hbar}\left(\frac{\pi}{6}\right)^{1/3}\rho_{atom}^{1/3}, \qquad \text{(Equation 2)}$$

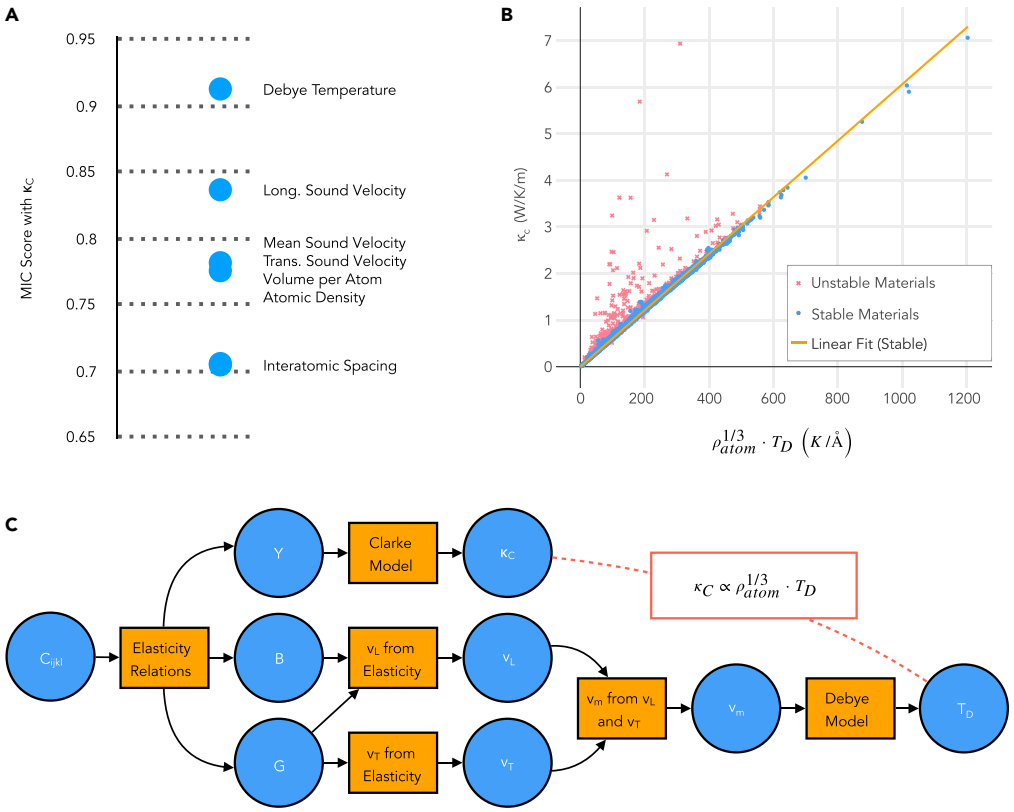


Figure 5. Example of Propnet Model Discovery

(A) Ranking of the top seven scalar properties with the highest MIC score when correlated with Clarke thermal conductivity.

(B) Clarke thermal conductivity versus the product of the Debye temperature with the cube root of atomic density for a representative sample of dynamically stable materials (blue dots) with a least-squares linear fit to the full stable materials dataset, constrained to intersect the origin (orange line). Dynamically unstable materials (red crosses) are shown to demonstrate their poor adherence to the model.

(C) Schematic of the new relationship found by `propnet`. The `propnet`-discovered direct relationship described in the text is symbolized by the dashed, red line. Symbol key: Clarke thermal conductivity (κ_C), Debye temperature (T_D), atomic density (ρ_{atom}), mean/longitudinal/transverse sound velocity ($v_m/v_L/v_T$), Young's/bulk/shear modulus ($Y/B/G$), and elastic tensor (C_{ijkl})

where k_B is Boltzmann's constant, $\hbar$ is the reduced Planck's constant, and ρ_{atom} is the atomic density. When these two properties are plotted against one another, we note that the relationship is almost linear, demonstrating a weak dependence on the atomic density (Figure 5).

This nearly linear relationship between Debye temperature and Clarke thermal conductivity is not included explicitly as a model in `propnet`. However, through correlation analysis, `propnet` is able to uncover the strong relationship, suggesting a new, shorter calculation pathway between the two properties. While this specific example is explained intuitively, it demonstrates the ability of `propnet` to identify unknown relationships between larger and more disparate sets of materials properties.

A subset of materials deviates significantly from the nominally linear relationship. These materials are not anomalous in terms of their atomic density, but instead exhibit a dynamically unstable structure, which results in an elastic tensor with one or more negative eigenvalues (Figure 5, red crosses). The models for Debye temperature and Clarke thermal conductivity are formulated in expectation of a positive definite tensor signifying a dynamically stable structure; hence, these data are spurious.

Uncertainty in Materials Data

Focusing on a small subset of mechanical properties, we explore in detail the ability of `propnet` to gauge the uncertainty within a materials dataset. As an example, we consider Vickers hardness: an engineering property of materials that is most closely related to the yield strength.[31] A rough estimate of Vickers hardness can be obtained from the elastic tensor, which can serve (1) as a first approximation of a material's hardness and wear and (2) as a screening metric for superhard materials.[32–37] From the available data in the Materials Project, Vickers hardness can be estimated by `propnet` via numerous distinct pathways (Figure 6). Each separate path involves different models that, when evaluated in concert, result in the Vickers hardness. Thus, for a single material, many different Vickers hardness values can be obtained: one value per calculation path. When multiple values are provided for various properties, all combinations of values are considered in order to form a complete collection of input sets. For example, if `propnet` were presented with two elastic tensors—one experimental and one from DFT—there would be a cascade of new possible values for the Vickers hardness. For each calculation pathway presented above, `propnet` uses any combination of values derived from experiment or theory. As such, this framework automatically generates the ensemble of possible values for any material property, consistent with any combination of input data and `propnet` pathways.

To illustrate this, we calculate values for Vickers hardness using another material, wurtzite GaN. As a result, we obtain an ensemble of 128 calculated values, each originating from a unique pathway encoded in `propnet`. Some paths start with the elastic tensor and use different polycrystalline averaging schemes[38] to obtain the elastic moduli; others derive the elastic moduli using a machine-learning model[39] based on structural and energetic properties. Selecting one particular property and examining its spread therefore gives a practical metric for gauging the uncertainty in that value. For our ensemble of Vickers hardness values, we observe a mean value of 11.9 GPa with a standard deviation of 4.25 GPa (Figure 6). The Vickers hardness of wurtzite GaN has been experimentally measured to be $\approx$10.0 GPa.[40] While this value was not output exactly by any one pathway, it is near the middle of the predicted distribution.

Uncertainty Quantification and Combination of Models

In the case that several physical or empirical models connect two different materials properties, `propnet` is equipped with the ability to benchmark the models by their accuracy. By performing a linear, least-squares prediction, `propnet` can also combine different models to optimize a connection between two properties.

As an example, we consider the case of estimating the refractive index from the band gap. In our current version of `propnet`, there are five different models connecting the two properties: an original model published by Moss[24] and four derivative models published by Hervé and Vandamme,[28] Gupta and Ravindra,[25] Reddy and Nazeer Ahammed,[27] and Reddy et al.[26] We use `propnet` to estimate the refractive indices for a subset of Materials Project materials, filtering for those with experimental values of both the band gap and the refractive index at visible and infrared frequencies (see Experimental Procedures: Model Benchmarking). Comparing the predictions from `propnet` with both the visible and infrared experimental datasets, we observe considerable differences between the predicted and measured values, particularly in materials with refractive indices less than $\approx$2.5 (Figure 7, blue circles). For these materials, the Ravindra model performs poorly, giving relatively high root-mean-square error (RMSE) values for both the visible and infrared datasets (Table 1).

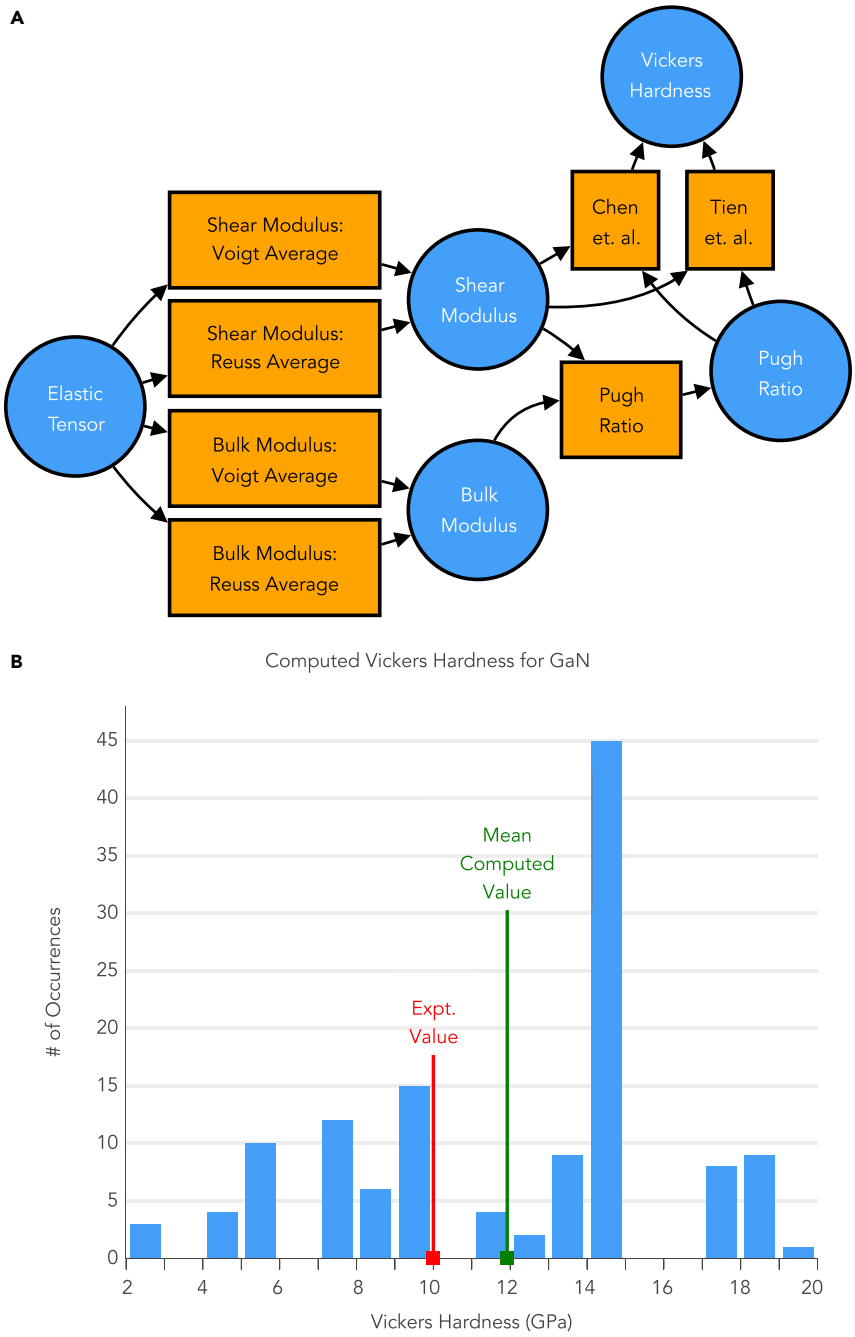


Figure 6. Example of Propnet Calculation of Vickers Hardness

(A) Schematic illustrating a selection of possible calculation pathways to reveal the Vickers hardness from the elastic tensor.

(B) Histogram of the derived Vickers hardness values for wurtzite GaN from the DFT-calculated elastic tensor, structural, and energy data. The approximate experimental value from Yonenaga and Suzuki[40] is depicted by the red line and the mean of the computed values is depicted by the green line.

In contrast, the Reddy/Anjaneyulu model performed well over both datasets, and the Hervé/Vandamme and Moss models had comparatively low RMSE for the visible and infrared datasets, respectively.

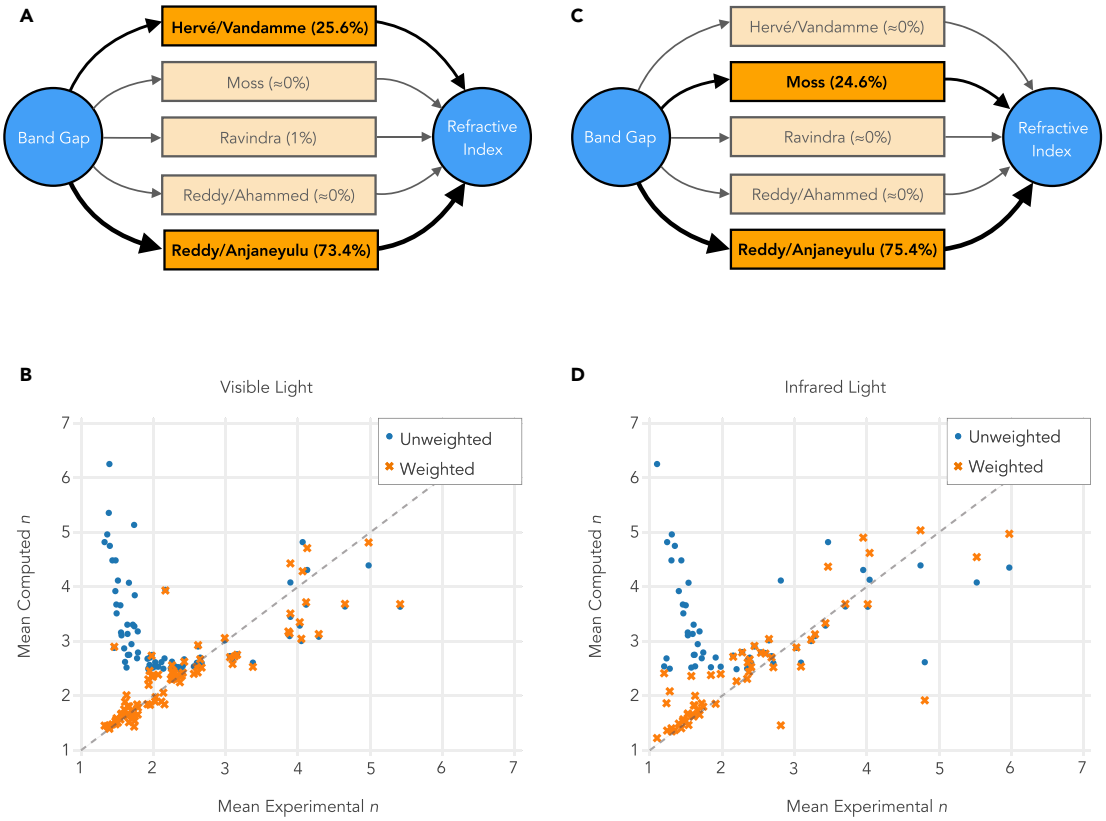


Figure 7. Improving Accuracy of Property Prediction by Model Weighting

Illustration of `propnet`'s ability to benchmark models against a set of known values. On top, schematics depict the ideal weighting of several refractive index (*n*) models when calculating the mean *n* value for materials using `propnet`; weights were benchmarked with experimental *n* values measured using (A) visible (λ = 589.3 nm) and (C) infrared light (λ = 10.0 μm). On the bottom, the plots compare mean `propnet`-computed *n* values for a set of materials with their respective experimental *n* values, measured using (B) visible and (D) infrared light. The mean `propnet`-computed *n* is shown with equal model weighting (blue circles) and benchmarked weights (orange crosses).

The RMSE information is inherently important and can be visually inspected. However, it is even more useful to analyze the deviation between each model and the collected experimental data to arrive at an improved relationship between the band gap and the refractive index. By default, `propnet` aggregates property values of the same type using an unweighted mean, i.e., values derived from each model are weighted equally. This protocol is not ideal in situations where models only work well over a range of values or material types. For example, the Ravindra model is known to give poor, non-physical results for materials with low refractive indices, and the Reddy/Ahammed model cannot be evaluated for materials with small band gaps.[29] Additionally, the Hervé/Vandamme model performs better with refractive indices measured with visible light because the empirical model was originally fit using refractive index data collected in the visible region. Likewise, the Moss model was based primarily on refractive index data collected using infrared light.

To address these disparities in performance, `propnet` implements a benchmarking protocol, which generates an improved weighting of models for aggregation by fitting to experimental data using a constrained linear least-squares fit (see Model Benchmarking for details). The various biases of the models became apparent after running the benchmarking procedure with the two datasets: `propnet` assigns large weights to the Reddy/Anjaneyulu and Hervé/Vandamme models on the visible

Table 1. List of Models in `propnet` Used to Calculate Refractive Index from Band Gap with Their Root-Mean-Square Errors and the Benchmarked Weighted Percentages for the Visible and Infrared Experimental Datasets

Model	RMSE		% Weight	
	Visible	Infrared	Visible	Infrared
Hervé/Vandamme	0.52	1.15	25.6%	≈0%
Moss	0.54	0.93	≈0%	24.6%
Ravindra	6.84	7.13	1%	≈0%
Reddy/Ahammed*	0.78	0.98	≈0%	≈0%
Reddy/Anjaneyulu	0.48	0.90	73.4%	75.4%
Unweighted mean	1.50	1.80		
Weighted mean	0.46	0.89		

The last two rows show root-mean-square errors (RMSE) values for unweighted and weighted means of refractive indices obtained from these models. See de Jong et al.[29] for mathematical forms of each model.

*The Reddy/Ahammed model could not be evaluated for materials with band gaps less than 0.365 eV, causing them to be excluded from the RMSE calculation, producing an artificially low model RMSE.

dataset; propnet assigns large weights to the Reddy/Anjaneyulu and Moss models on the infrared dataset; and in all cases, `propnet` minimizes the contribution of the Ravindra model. With these benchmarked weights, the improvement in prediction can be readily seen (Figure 7, orange crosses) and the RMSE for the weighted mean drops below that of the unweighted mean and any of the individual models (Table 1). The results with a linear model are promising and highly interpretable, but one could also explore more complex ensemble methods, such as a random forest prediction.[41] With the implementation of improved averaging schemes, future versions of `propnet` could employ fitted relationships when aggregating property values during graph evaluation to reduce the propagation of errors during calculations of related properties.

Conclusion

In this work, we demonstrate a programmatic framework for systematically codifying, connecting, and analyzing materials properties. To our knowledge, no such framework exists today, and we anticipate a rapidly growing need as the field of materials informatics expands. To illustrate the features of `propnet`, we exemplify (1) the rapid increase in derived, available materials properties, (2) the facilitated examination of correlations between property sets, (3) uncertainty quantification, and (4) the ability of `propnet` to derive model weightings in response to a benchmarking dataset.

The currently implemented set of models is limited to mainly fundamental mechanical or dielectric quantities. However, the demonstrated behavior can be extended arbitrarily with the use of customized models and materials properties. Because `propnet` is readily extensible, this software package leverages not only simple analytical models but also combines previously coded algorithms for deriving materials properties (e.g., `pymatgen`)[42] and machine-learning models (e.g., `gbml`).[39] As such, `propnet` is able to combine together vastly different codebases to create a cohesive amalgamation of knowledge in a single resource.

In combination with a user's own data, we envision `propnet` to be a powerful tool for combining and augmenting many sources of materials data. The current version of `propnet` includes an adapter to automatically access data from the Materials

Project. In future versions, we anticipate the development of additional tools to access and combine data from other open-access resources in the rapidly growing domain of materials informatics.

EXPERIMENTAL PROCEDURES

The Python codebase for `propnet` is available under an open-source license and can be found at https://github.com/materialsintelligence/propnet. Please see the repository for details on usage.

Design Philosophy of Code

Of principal importance in the design of `propnet` is ensuring that materials property models can be faithfully represented programmatically. Eschewing blind application of simple formulas, `propnet` employs a flexible object-oriented approach that accommodates the encoding of constraints on model evaluation by means of method override. This fundamentally ensures that a result cannot be derived unless all required conditions are met for the model to be valid. Analyzing, for example, the Wiedemann-Franz-Lorenz law for metallic thermal conductivity, `propnet` will not generate an output unless it knows that the material's band gap is zero. However, `propnet` may proceed with a non-metal if a value for the Lorenz number is given as input.

To emphasize community accessibility, many models can be encoded using a simplified text file (YAML) format with representative equations directly specified in the file. An example of such a file can be found in Supplemental Information. For users with Python knowledge, more complex models can be created using a Python module template. To facilitate broad adoption, `propnet` includes automatic unit handling and conversion, utilizing the `pint` Python package.[43]

`propnet` employs a custom graph traversal algorithm to generate all possible outputs from a given input set. This process, termed a self-consistent evaluation scheme, is used to deterministically augment datasets while avoiding infinite loops. Fundamentally, the set of connected materials properties and property relationships may be cyclic. Thus, as outputs are generated from property relationships, these outputs may correspond to new inputs for other property relationships. In the general case, `propnet` must continuously revisit models and account for these new inputs, thereby generating additional outputs.

To prevent an infinite loop in the case of a cyclic graph, each output from a model is tagged as originating from that model. If the input set that generated the output had elements derived from other models, the output is tagged as originating from these models as well. Maintaining a set of connections, an output can be generated from a model and an input set if, and only if, none of the input set values have been derived from the same model. There exist many other potential schemes for avoiding infinite loops when evaluating models, and the precise halting criterion is generally flexible. In particular, this approach was selected to generate all intuitive outputs while minimizing redundant calculations of interdependent properties.

Emphasizing abstraction, the graph of materials properties and property relationships exists independent of any materials data. Models and properties taken together compose the final property network graph, representing the means by which materials properties are calculated. Conforming to this model, we take

collections of materials property data to exist separate from the graph. Thus, the action of the graph is solely to augment these sets of data with additional values.

Model Benchmarking

The experimental band gap and refractive index datasets used for model benchmarking are available in Supplemental Information with references for individual values used.

Experimental refractive index data were retrieved from the RefractiveIndex.info database,[44] which tabulates refractive indices as a function of the wavelength of light used to measure the refractive index (i.e., dispersion) from one or more reference sources. Data explicitly labeled as simulation or model data was excluded. Materials were selected based on the availability of data measured at the wavelength of choice: visible light, 589.3 nm (sodium D lines) or infrared light, 10.0 μm. For each reference source containing relevant data about a material, specific wavelength-dependent refractive index values were obtained by evaluating the empirical, mathematical form of the dispersion function or by linear interpolation of the discrete data, depending on the format provided. If multiple reference sources were available for a given material, a subset of values was created from the refractive indices within one standard deviation of the superset mean. The refractive index for the material was then taken as the mean value of the subset. Materials with mean refractive indices ≤ 1 were excluded.

Experimental band-gap data were retrieved from the Citrination database[6] by chemical formula search for the materials in each refractive index dataset. If no experimental band gap was available, the material was excluded from the dataset. If multiple band gaps were available for a given material, the subset mean approach was applied as described above.

Benchmarking of weights was done by minimizing the sum of squared errors (least squares) using the implementation of the Trust-Region Constrained Algorithm[45] as implemented in `scipy`.[46] Weights were constrained to values between 0 and 1 where the sum of all weights must equal 1. If a refractive index generated by a given model was unavailable for a material, the weights for the remaining models were scaled such that their sum was 1 for that material.

ACKNOWLEDGMENTS

We acknowledge the Toyota Research Institute for funding the development of this public codebase. Connections with the Materials Project were additionally supported by the Materials Project Program (grant no. KC23MP) through the DOE Office of Basic Energy Sciences, Materials Sciences, and Engineering Division, under contract DE-AC02-05CH11231. We acknowledge John Dagdelen, Leigh Weston, and Joseph Palakapilly for their input at the start of the project and John Dagdelen for the propnet name. We acknowledge Donny Winston and Patrick Huck for their input related to the website and database building.

AUTHOR CONTRIBUTIONS

D.M. established code architecture, focusing on graph traversal algorithms as well as infrastructural design, wrote of the text of the paper, and created several figures. M.K.H. established architectural design, wrote the initial implementation of the code and several models, maintained the code repository, shared coding responsibilities throughout the project, and created the web app interface. J.H.M. focused

on programmatic development of a user-friendly model and property specification format as well as infrastructural design. C.M.L. finalized the package, contributing to infrastructural design and implementing tools for external data access, data storage, and correlation analysis, additionally contributing to the text and figures. S.D. contributed to discussion and implementation of infrastructure. S.D. and V.T. contributed a number of models and properties to the core set. K.A.P. and A.J. conceptualized and directed the work.

DECLARATION OF INTERESTS

Code developed under the auspices of the Toyota Research Institute. Code is public and free for use. No personal interests conflict with the mission outlined in this paper.

REFERENCES

1. Nicole, L., Laberty-Robert, C., Rozes, L., and Sanchez, C. (2014). Hybrid materials science: a promised land for the integrative design of multifunctional materials. Nanoscale *6*, 6267–6292.
2. Belsky, A., Hellenbrandt, M., Karen, V.L., and Luksch, P. (2002). New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. Acta Crystallogr. Sect. B Struct. Sci. *58*, 364–369.
3. Xu, Y., Yamazaki, M., and Villars, P. (2011). Inorganic materials database for exploring the nature of material. Jpn. J. Appl. Phys. *50*, 11RH02.
4. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. (2013). Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. *1*, 011002.
5. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM *65*, 1501–1509.
6. Mullin, R. (2017). Citrine informatics. Chem. Eng. News *95*, 34.
7. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555.
8. Tran, K., Palizhati, A., Back, S., and Ulissi, Z.W. (2018). Dynamic workflows for routine materials discovery in surface science. J. Chem. Inf. Model. *58*, 2392–2400.
9. P. Linstrom, and W. Mallard, eds. (2019). NIST Chemistry WebBook, NIST Standard Reference Database Number 69 (National Institute of Standards and Technology). https://doi.org/10.18434/T4D303.
10. Zhang, Y., and Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. NPJ Comput. Mater. *4*, 25.
11. Fermi, E. (1928). Eine statistische methode zur bestimmung einiger eigenschaften des atoms und ihre anwendung auf die theorie des periodischen systems der elemente. Z. Phys. *48*, 73–79.
12. Hohenberg, P., and Kohn, W. (1964). Inhomogeneous electron gas. Phys. Rev. *136*, B864–B871.
13. Levy, M. (1982). Electron densities in search of Hamiltonians. Phys. Rev. A *26*, 1200–1208.
14. Jones, R.O., and Gunnarsson, O. (1989). The density functional formalism, its applications and prospects. Rev. Mod. Phys. *61*, 689–746.
15. Perdew, J.P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. Phys. Rev. Lett. *77*, 3865–3868.
16. Anisimov, V.I., Zaanen, J., and Andersen, O.K. (1991). Band theory and Mott insulators: Hubbard U instead of Stoner I. Phys. Rev. B *44*, 943–954.
17. de Jong, M., Chen, W., Angsten, T., Jain, A., Notestine, R., Gamst, A., Sluiter, M., Krishna Ande, C., van der Zwaag, S., Plata, J.J., et al. (2015). Charting the complete elastic properties of inorganic crystalline compounds. Sci. Data *2*, 150009.
18. de Jong, M., Chen, W., Geerlings, H., Asta, M., and Persson, K.A. (2015). A database to enable discovery and design of piezoelectric materials. Sci. Data *2*, 150053.
19. Tran, R., Xu, Z., Radhakrishnan, B., Winston, D., Sun, W., Persson, K.A., and Ong, S.P. (2016). Surface energies of elemental crystals. Sci. Data *3*, 160080.
20. Petousis, I., Mrdjenovich, D., Ballouz, E., Liu, M., Winston, D., Chen, W., Graf, T., Schladt, T.D., Persson, K.A., and Prinz, F.B. (2017). High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. Sci. Data *4*, 160134.
21. Petretto, G., Dwaraknath, S., Miranda, H.P.C., Winston, D., Giantomassi, M., van Setten, M.J., Gonze, X., Persson, K.A., Hautier, G., and Rignanese, G.M. (2018). High-throughput density-functional perturbation theory phonons for inorganic materials. Sci. Data *5*, 180065.
22. Luo, B., Deng, Y., Wang, Y., Tan, M., Cao, L., and Zhu, W. (2012). Fabrication and growth mechanism of zinc blende and wurtzite CdTe nanowire arrays with different photoelectric properties. CrystEngComm *14*, 7922.
23. Kinney, J.B., and Atwal, G.S. (2014). Equitability, mutual information, and the maximal information coefficient. Proc. Natl. Acad. Sci. U S A *111*, 3354–3359.
24. Moss, T.S. (1951). Photoconductivity in the elements. Proc. Phys. Soc. Lond. Sect. A *64*, 590–591.
25. Gupta, V.P., and Ravindra, N.M. (1980). Comments on the Moss formula. Phys. Status Solidi B *100*, 715–719.
26. Reddy, R., Anjaneyulu, S., and Sarma, C. (1993). Relationship between energy gap, refractive index, bond energy and the Szigeti charge in polyatomic binary compounds and semiconductors. J. Phys. Chem. Sol. *54*, 635–637.
27. Reddy, R., and Nazeer Ahammed, Y. (1995). A study on the Moss relation. Infrared Phys. Technol. *36*, 825–830.
28. Hervé, P., and Vandamme, L. (1994). General relation between refractive index and energy gap in semiconductors. Infrared Phys. Technol. *35*, 609–615.
29. Ravindra, N., Ganapathy, P., and Choi, J. (2007). Energy gap-refractive index relations in semiconductors—an overview. Infrared Phys. Technol. *50*, 21–29.

30. Clarke, D.R. (2003). Materials selection guidelines for low thermal conductivity thermal barrier coatings. Surf. Coat. Technol. *163-164*, 67–74.

31. Tiryakioğlu, M. (2015). On the relationship between Vickers hardness and yield stress in Al–Zn–Mg–Cu alloys. Mater. Sci. Eng. A *633*, 17–19.

32. Kaner, R.B., Gilman, J.J., and Tolbert, S.H. (2005). Designing superhard materials. Science *308*, 1268–1269.

33. Oganov, A.R., and Lyakhov, A.O. (2010). Towards the theory of hardness of materials. J. Superhard Mater. *32*, 143–147.

34. Gao, F.M., and Gao, L.H. (2010). Microscopic models of hardness. J. Superhard Mater. *32*, 148–166.

35. Chen, X.Q., Niu, H., Li, D., and Li, Y. (2011). Modeling hardness of polycrystalline materials and bulk metallic glasses. Intermetallics *19*, 1275–1281.

36. Mansouri Tehrani, A., and Brgoch, J. (2019). Hard and superhard materials: a computational perspective. J. Solid State Chem. *271*, 47–58.

37. Avery, P., Wang, X., Proserpio, D.M., Toher, C., Oses, C., Gossett, E., Curtarolo, S., and Zurek, E. (2019). Predicting superhard materials via a machine learning informed evolutionary structure search. http://arxiv.org/abs/1906.05886.

38. Hill, R. (1952). The elastic behaviour of a crystalline aggregate. Proc. Phys. Soc. Lond. Sect. A *65*, 349–354.

39. de Jong, M., Chen, W., Notestine, R., Persson, K., Ceder, G., Jain, A., Asta, M., and Gamst, A. (2016). A statistical learning framework for materials science: application to elastic moduli of *k*-nary inorganic polycrystalline compounds. Sci. Rep. *6*, 34256.

40. Yonenaga, I., and Suzuki, T. (2002). Indentation hardnesses of semiconductors and a scaling rule. Philos. Mag. Lett. *82*, 535–542.

41. Breiman, L. (2001). Machine Learning (Kluwer Academic Publishers).

42. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput. Mater. Sci. *68*, 314–319.

43. Grecco, H. (2019). pint: operate and manipulate physical quantities in python. https://github.com/hgrecco/pint.

44. Polyanskiy, M.N. (2019). Refractive index database. https://refractiveindex.info.

45. Lalee, M., Nocedal, J., and Plantenga, T. (1998). On the implementation of an algorithm for large-scale equality constrained optimization. SIAM J. Optim. *8*, 682–706.

46. Jones, E., Oliphant, T., and Peterson, P. (2001). SciPy: open source scientific tools for Python. http://www.scipy.org.